Proyecto entorno al problema de estimación de provisiones en la industria aseguradora empleando el método Chain-ladder

Santiago Prieto Betancur

Índice

1 FASE DE COMPRENSIÓN DEL NEGOCIO/PROBLEMA

En el marco de la metodología de desarrollo de proyectos de CRISP-DM (Cross-Industry Standard Process for Data Mining) es necesario establecer en un primer momento el entendimiento del negocio o problema a tratar el cual empleará herramientas del análisis de datos.

De esta manera, en el problema de estimación de provisiones (reservas por pérdidas) en la industria aseguradora el objetivo principal es destinar un fondo de reservas o de aprovisionamientos de una cartera con el fin de liquidar aquellos productos de seguros reportados por los clientes después de un determinado siniestro. Por lo tanto, en el problema de estimación de provisiones el intento es estimar las reclamaciones de seguros que debe atender una empresa aseguradora pero que aun no han sido reportados y proyectar adicionalmente las ultimas cantidades de perdida en las cuales podría incurrir una compañía atendiendo estos siniestros. Adicionalmente, es de gran importancia entender que este problema al tratarse de una cuestión de naturaleza predictiva implica que una de las suposiciones subyacentes en la solución de esta incógnita es que los desarrollos de patrones que explican las perdidas históricas en la estimación de estas provisiones son un indicador para desarrollar patrones para perdidas futuras.

1.1 OBJETIVOS DEL NEGOCIO/PROBLEMA

A medida que el mercado asegurador se expande y más compañías incursionan en este negocio es importante desarrollar e implementar mecanismos que hagan rentable la oferta de seguros accediendo a un amplio número de clientes y a la par manteniendo una oferta de productos competitivos en el mercado. Es por esto, que se hace indispensable desarrollar un plan de trabajo en el cual se cumplan los siguientes objetivos:

- Recopilar, organizar y analizar datos relacionados con las pólizas a pagar después del reporte de un siniestro por los clientes de una compañía aseguradora.
- Determinar patrones o modelos que permitan reservar los montos adecuados de aprovisionamiento para los gastos operativos en una compañía aseguradora.
- Optimizar las reservas de cartera apresadas con el fin de mantener un flujo de capital mayor en inversiones o financiación de la compañía aseguradora y a su vez estimar el capital necesario para cubrir la liquidación de las pólizas reportadas.

De esta manera, para la consecución de este proyecto es necesario de la cooperación de distintos departamentos dentro de la estructura organizacional de una compañía de seguros

como lo son, por lo general, el comité de manejo del riesgo, el comité de inversión, el comité de auditoría y los asesores en el área de actuaria y IT (Information Technology) ya que de esta manera es posible articular mejor la información y a su vez tener acceso a datos de interés en el estudio y permitir igualmente su correcto procesamiento y análisis.

1.2 EVALUACIÓN DE LA SITUACIÓN

Para este proyecto es necesario disponer de una base de datos que permita determinar si el modelo predictivo a desarrollar brinda buenos resultados al momento de fijar las reservas de provisiones en una compañía aseguradora. Es por esto, que en el trabajo a desarrollar se empleará la base de datos CAS Loss Reserve la cual fue construida a partir de la base de datos Schedule P – Analysis of Losses and Loss Expenses in the National Association of Insurance Commissioners (NAIC).

Esta base de datos contiene información de los reportes hechos por las principales líneas personales y comerciales de todas las aseguradoras de daños y perjuicios que están registradas en los Estados Unidos. En la base de datos existen seis lineas de aseguramiento las cuales son:

- private passenger auto liability/medical
- commercial auto/truck liability/medical
- workers' compensation
- medical malpractice – claims made
- other liability – occurrence
- product liability – occurrence

Para este trabajo solo se dispondrá de la base de datos relacionada con los seguros de las malas practicas médicas (medical malpractice – claims made) con el fin de explorar solo un caso de aseguramiento. Es importante esclarecer desde un principio los limitantes de este tipo de proyectos y las suposiciones en los cuales se cimienta. El problema de estimación de provisiones posee el limitante o desventaja de que solo es preciso cuando los patrones del desarrollo de perdidas del pasado continúan funcionando en el futuro, por esta razón cuando existen cambios en las operaciones de aseguramiento como los cambios en los tiempos de liquidación de las reclamaciones o los cambios en las prácticas de reserva de provisiones puede suceder que los métodos a desarrollar no produzcan estimaciones precisas de aprovisionamiento si no se hacen los respectivos ajustes al modelo a desarrollar por lo cual se tiene que este tipo de problemas son muy sensibles a los cambios y pueden ser impropios para ciertas lineas de negocio muy volátiles.

1.3 OBJETIVOS EN LA CIENCIA DE DATOS

Una vez establecidos los objetivos del problema es necesario identificar como estos objetivos se ven traducidos en las labores de la ciencia de datos, es decir determinar cual es el propósito del análisis de datos que se desarrollara durante el proyecto. De esta manera, en el tratamiento de los datos se fijan los siguientes objetivos:

- Identificar los patrones que permitan estimar la reserva de provisiones en determinados tipos de seguros a partir de la base de datos histórica CAS Loss Reserve expuesta anteriormente. (Modelo de forecasting)
- Establecer medidas de desempeño que permitan valorar el modelo a desarrollar en la base de datos del proyecto y que sean compatibles con los triángulos de perdida que provee la base de datos.

1.4 PLAN DE PROYECTO

Para la consecución de los objetivos de este proyecto en primer lugar es necesario iniciar una fase de entendimiento de los datos para comprender la distribución y la información que recopila la base de datos. Posteriormente, dado que los datos ya se encuentran ordenados, se inicia la fase de modelado en donde se busca poder determinar las reservas de provisiones a partir de registros históricos consignados en los triángulos de perdida usando el método de Chain-ladder. Una vez establecido el modelo se procede a fijar una medida de desempeño adecuada que permita evaluar la capacidad predictiva de nuestro modelo con los datos de evaluación extraídos de la base de datos. Una vez completas todas las fases y después de un proceso de depuración y corrección de los métodos empleados y los resultados obtenidos se procede a la fase de despliegue del modelo.

2 FASE DE ENTENDIMIENTO DE LOS DATOS

En esta parte del proyecto se recopila toda la información relacionada con los datos disponibles para realizar este proyecto en analítica de datos. El objetivo de esta parte del proyecto es inspeccionar los datos disponibles con el fin de evaluar la calidad de los datos y por consiguiente detectar y evitar aquellos imprevistos que podrían ocasionar problemas inesperados.

2.1 RECOLECCIÓN DE LOS DATOS INICIALES

Esta parte del proyecto, en este caso particular, no representa un problema considerable ya que se está usando la base de datos CAS Loss Reserve la cual previamente ya fue organizada

de tal manera que podamos acceder a lo datos ya recopilados. Como se expreso en fases previas de este proyecto la base de datos tomada para analizar es la de malas practicas médicas y puntualmente se analizará la variable relacionada con CumPaidLoss_F2 (Pérdidas pagadas acumuladas y gastos asignados al final del año).

2.2 DESCRIPCIÓN DE LOS DATOS

La descripción inicial, análisis de los datos y toda la estructura tabular de la base de datos se puede encontrar en el siguiente cuaderno de Colab https://github.com/SantiagoUNAL/ML-Applications-in-Actuarial-Sciences/blob/9ea1af30cb8864273ce246685e7bfc60aad674ba/Preparaci%C3%B3n_de_los_datos_Proyecto_1_(Santiago_Prieto_Betancur).ipynb. Sin embargo, es importante puntualizar esta descripción en las variables que se van a analizar en este proyecto. A lo largo del siguiente trabajo se dispondrá de 34 datos donde cada uno representa un triangulo de perdida correspondiente a una aseguradora o a un conglomerado de aseguradoras afiliadas que recopila la información asociada a la variable CumPaidLoss_F2 (Pérdidas pagadas acumuladas y gastos asignados al final del año) en esta base de datos se tienen registrados los reportes atendidos en los años de accidentalidad de 1988 a 1997 (10 años) con 10 años de retraso en el desarrollo.

2.3 EXPLORACIÓN DE LOS DATOS Y CALIDAD DE LOS DATOS

En primer lugar, observemos que la distribución de los datos en la variable CumPaidLoss_F2 muestra que existen una gran cantidad de valores nulos en los cuadrados de perdidas asociados a cada entidad aseguradora lo cual deja como precedente que los datos en su mayoría son esparsos como lo muestra el siguiente histograma

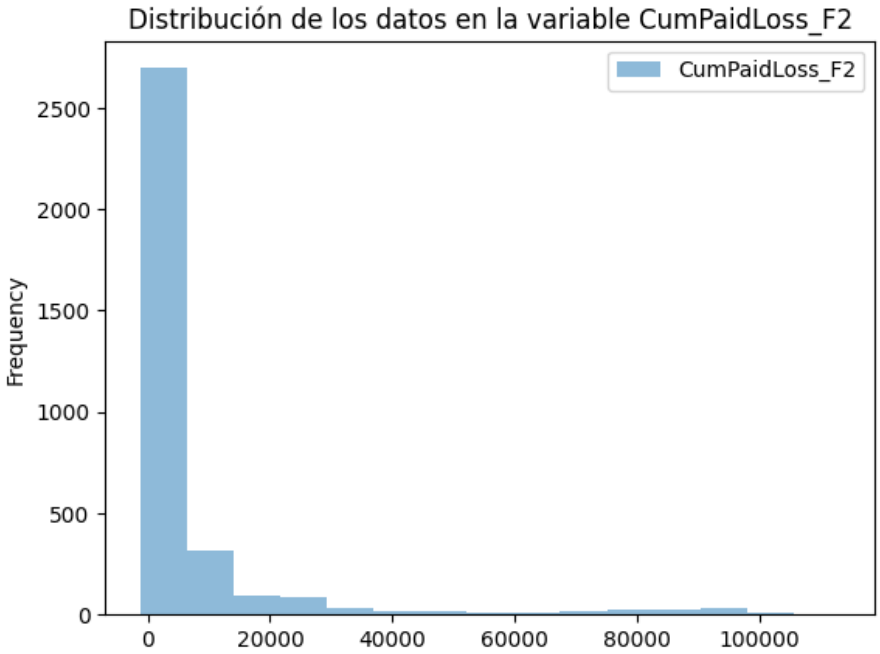

FIGURE 1: HISTOGRAMA DE LOS DATOS EN LA VARIABLE CUMPAIDLOSS_F2

Del análisis realizado en el cuaderno de Colab encontramos que la media de los cuadrados asociados a cada aseguradora en la variable CumPaidLoss_F2 es

$$\overline{X} = \begin{bmatrix} 174{,}529 & 1389{,}941 & 2795{,}706 & 3871{,}059 & 4856{,}382 & 5351{,}088 & 5835{,}147 & 6060{,}824 & 6275{,}706 & 6389{,}382 \\ 277{,}441 & 1525{,}147 & 2911{,}059 & 4387{,}324 & 5164{,}353 & 5811{,}471 & 6213{,}000 & 6411{,}353 & 6550{,}206 & 6601{,}618 \\ 352{,}824 & 1610{,}059 & 3498{,}941 & 4814{,}559 & 5599{,}735 & 6293{,}294 & 6623{,}500 & 6932{,}853 & 7107{,}794 & 7191{,}294 \\ 279{,}912 & 2159{,}412 & 4304{,}324 & 5860{,}647 & 7205{,}500 & 7656{,}853 & 8115{,}382 & 8400{,}853 & 8587{,}618 & 8657{,}588 \\ 367{,}029 & 2300{,}353 & 4629{,}412 & 6175{,}265 & 7177{,}000 & 7853{,}147 & 8338{,}676 & 8582{,}265 & 8846{,}265 & 8981{,}618 \\ 536{,}147 & 2667{,}941 & 4891{,}912 & 6702{,}676 & 8124{,}559 & 9014{,}559 & 9433{,}500 & 9838{,}294 & 10004{,}735 & 10189{,}029 \\ 439{,}765 & 2773{,}618 & 5487{,}559 & 7424{,}971 & 8740{,}206 & 9593{,}529 & 10091{,}912 & 10366{,}088 & 10696{,}265 & 10856{,}529 \\ 529{,}265 & 3240{,}618 & 6153{,}588 & 8081{,}529 & 9474{,}029 & 10462{,}029 & 10881{,}382 & 11285{,}912 & 11541{,}882 & 11747{,}824 \\ 599{,}706 & 3161{,}000 & 5974{,}618 & 7903{,}559 & 9479{,}676 & 10445{,}735 & 10992{,}324 & 11282{,}529 & 11492{,}971 & 11607{,}059 \\ 598{,}853 & 3312{,}824 & 6747{,}471 & 9183{,}824 & 10668{,}735 & 11645{,}735 & 12198{,}765 & 12670{,}235 & 12859{,}647 & 13152{,}882 \end{bmatrix}$$

Y la matriz de varianzas de las entradas de los cuadrados es

$$\Sigma^2 = \begin{bmatrix} 895{,}898 & 56821{,}661 & 229881{,}511 & 440738{,}130 & 693660{,}281 & 842180{,}744 & 1001439{,}447 & 1080399{,}466 & 1158367{,}186 & 1200711{,}966 \\ 1589{,}085 & 64436{,}333 & 230927{,}460 & 525347{,}370 & 730112{,}966 & 906436{,}050 & 1055297{,}059 & 1112554{,}156 & 1158551{,}773 & 1177610{,}150 \\ 3661{,}307 & 75299{,}654 & 351890{,}414 & 663200{,}321 & 897401{,}518 & 1111809{,}428 & 1232154{,}596 & 1348764{,}478 & 1417308{,}987 & 1448306{,}375 \\ 2304{,}429 & 118875{,}161 & 501995{,}121 & 859759{,}777 & 1264729{,}596 & 1438235{,}066 & 1633032{,}323 & 1760555{,}689 & 1846579{,}900 & 1879336{,}849 \\ 3665{,}581 & 134738{,}550 & 468100{,}183 & 805501{,}749 & 1120399{,}529 & 1199120{,}675 & 1255556{,}712 & 1346141{,}369 & 1315673{,}082 & 1369463{,}465 \\ 6922{,}578 & 171385{,}647 & 455861{,}007 & 767905{,}695 & 1239039{,}819 & 1527485{,}002 & 1710242{,}654 & 1896635{,}989 & 1976072{,}632 & 2065931{,}000 \\ 1004{,}059 & 132265{,}271 & 383840{,}625 & 508491{,}747 & 864396{,}271 & 1098193{,}226 & 1170657{,}872 & 998675{,}215 & 1115780{,}639 & 1174601{,}529 \\ 7272{,}711 & 144250{,}921 & 655965{,}872 & 1136624{,}645 & 1575965{,}606 & 2012056{,}692 & 2212065{,}167 & 2422711{,}011 & 2561294{,}949 & 2666621{,}274 \\ 8128{,}432 & 168284{,}235 & 530904{,}468 & 925504{,}417 & 1475837{,}422 & 1899679{,}829 & 1848773{,}941 & 1987078{,}346 & 2090129{,}533 & 2147086{,}505 \\ 8163{,}942 & 237027{,}030 & 343905{,}268 & 822073{,}619 & 1288849{,}864 & 1732398{,}877 & 1993967{,}095 & 2228856{,}515 & 2326904{,}303 & 2482858{,}298 \end{bmatrix}$$

De esto podemos ver que la variabilidad de las entradas en los triángulos aumenta a medida que los años de desarrollo de los siniestros aumenta lo cual es consecuente con el hecho de que se está analizando las pérdidas pagadas acumuladas y gastos asignados al final del año y estos tienden aumentar en el tiempo y a su vez incluye la variabilidad de los años de desarrollo previos, por lo cual es razonable que estos indicadores aumenten con el tiempo.

Del análisis exploratorio de los datos vemos que al estudiar la distancia de cada uno de los

cuadrados con respecto a la media usando la métrica de Frobenius que penaliza mayormente valores que difieren drásticamente, tenemos que existen datos notablemente distanciados del resto como se muestra a continuación

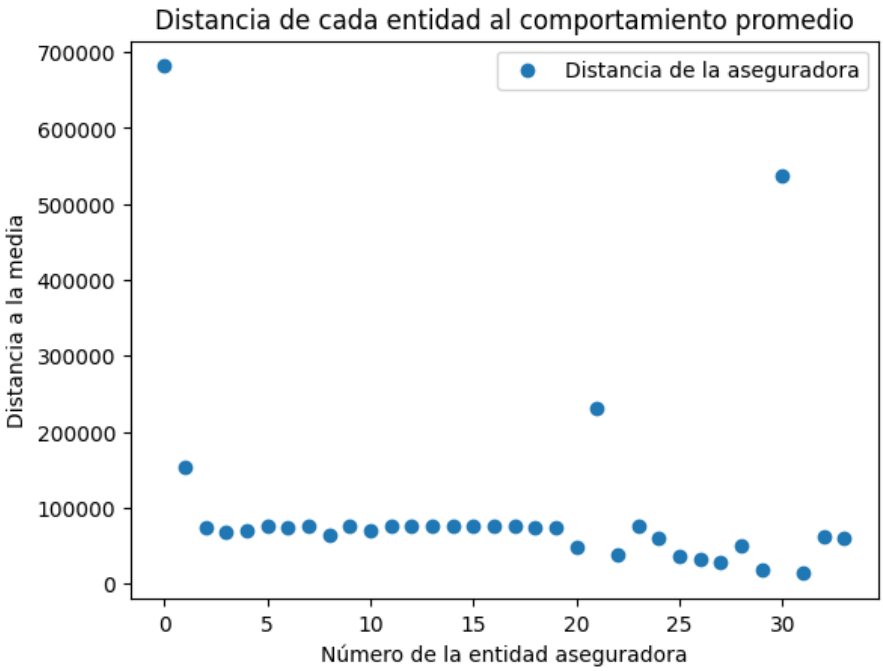

FIGURE 2: DISTANCIA DEL COMPORTAMIENTO DE CADA ASEGURADORA A LA MEDIA

De esta manera, los datos atípicos pertenecientes a esta base de datos corresponden a las siguientes aseguradoras:

- Scpie Indemnity Co
- Promutual Grp
- State Volunteer Mut Ins Co
- Physicians Recip Insurers

Con el objetivo de extraer los datos atípicos de esta base de datos, el cual posee un considerable número de matrices esparsas, miraremos la distancia de cada matriz de perdida con la matriz nula para identificar aquellas aseguradoras que no presentan muchas entradas en cero durante el periodo de análisis.

La gráfica de dispersión de las distancias de cada aseguradora es la siguiente:

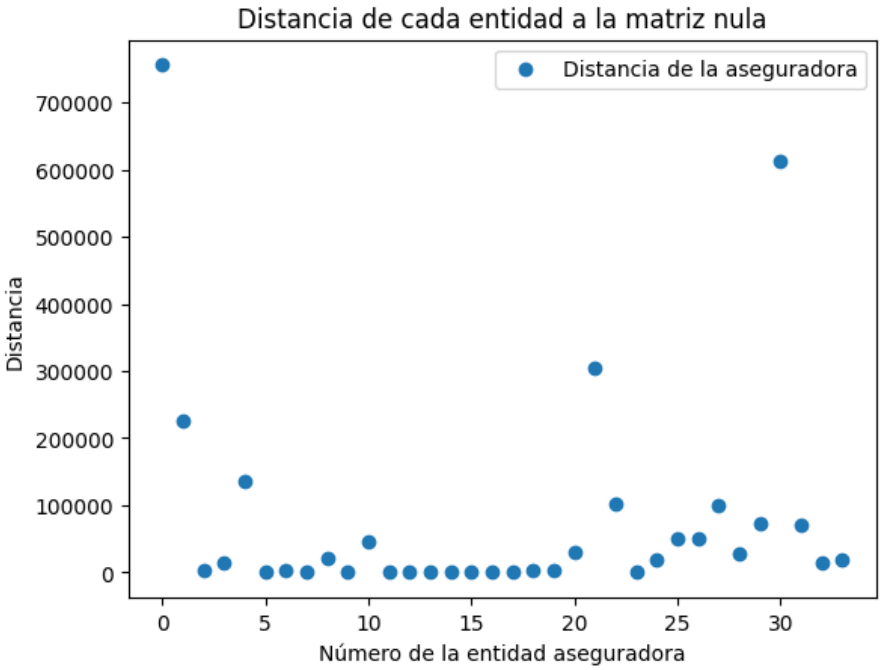

FIGURE 3: DISTANCIA DEL COMPORTAMIENTO DE CADA ASEGURADORA A LA MATRIZ NULA

De esta manera, las matrices que no son completamente nulas en esta base de datos corresponden a las siguientes aseguradoras:

- Scpie Indemnity Co
- Promutual Grp
- Nationwide Grp
- Markel Corp Grp
- Controlled Risk Ins Co Of VT Inc
- MCIC VT Inc RRG
- Texas Hospital Ins Exch
- State Volunteer Mut Ins Co
- MHA Ins Co
- National Guardian RRG Inc
- Preferred Professional Ins Co
- Medical Mut Ins Co Of ME
- Utah Medical Ins Assoc

- Seguros Triples Inc
- Dentists Ins Co
- Physicians Recip Insurers
- Louisiana Med Mut Ins Co
- Clinic Mut Ins Co RRG
- California Healthcare Ins Co Inc

3 FASE DE PREPARACIÓN DE LOS DATOS

En esta parte del proyecto destinamos un tiempo prudencial a la preparación, almacenamiento y estructuración de los datos de tal forma que sea posible servirse de los mismos en la fase de modelado. De esta manera, en el cuaderno de Colab se plantea todo un marco de trabajo en el cual se hace un tratamiento de los datos provistos con el fin de hacer posible su integración en los modelos subsiguientes.

Dentro del procesamiento de la base de datos se crearon herramientas de código que permitieran identificar a cada grupo de aseguradoras por su nombre o su código de registro. Además, se implementaron lineas de código que admiten estructurar los datos en los triángulos de perdida de forma tal que la manipulación de los datos sea más asequible para los objetivos del proyecto. Dentro de la organización provista a los datos se crearon tres diccionarios que almacenan los triángulos de perdida de cada grupo de aseguradoras en las siguientes categorías:

- El primer diccionario almacena en un formato de dataframe los triángulos superiores de perdida para todas las aseguradoras en la variable **CumPaidLoss_F2** sin considerar las estimaciones del triangulo inferior.
- El segundo diccionario almacena en un formato de dataframe los cuadrados de perdida para todas las aseguradoras en la variable **CumPaidLoss_F2**. Aquí se almacena toda la información del triangulo superior y del triangulo inferior.
- El tercer diccionario almacena en un formato de matriz los cuadrados de perdida para todas las aseguradoras en la variable **CumPaidLoss_F2**. Aquí se almacena toda la información del triangulo superior y del triangulo inferior como un array con el fin de alimentar los modelos a desarrollar posteriormente.

3.1 SELECCIÓN Y LIMPIEZA DE LOS DATOS

Después de realizar una análisis exploratorio de los datos y de cada uno de los triángulos de perdida de las aseguradoras se pudo identificar que existían grupos de aseguradoras que no tenían registros en la variable **CumPaidLoss_F2** antes de 1993 por lo que el comportamiento que describirían en el modelo a desarrollar no capturaría correctamente el patrón de aprovisionamiento. Esto pudo ocurrir principalmente porque la recopilación de la información para ese entonces no existía en esos grupos de aseguradoras o porque las compañías no existían todavía.

De esta manera, se depuro la base de datos y se extrajeron las aseguradoras que no tenían registros completos para que el modelo no fuese inconsistente en estos casos. Por lo tanto, después de los análisis previos se concluyo que el subconjunto de las aseguradoras que tienen registros completos durante el periodo de tiempo analizado son las siguientes:

- Scpie Indemnity Co
- Promutual Grp
- Markel Corp Grp
- Texas Hospital Ins Exch
- State Volunteer Mut Ins Co
- MHA Ins Co
- Preferred Professional Ins Co
- Utah Medical Ins Assoc
- Seguros Triples Inc
- Dentists Ins Co
- Physicians Recip Insurers
- Clinic Mut Ins Co RRG

Con la preselección de los datos anteriores es posible continuar con el desarrollo y evaluación de los modelos para el proyecto de ciencia de datos.

Es importante realizar una observación sobre la cantidad de datos disponibles. Dado que se tienen 12 datos para entrenar los modelos de aprovisionamiento, es necesario estructurar los datos bajo una estrategia de validación cruzada ya que no es adecuado dividir los datos en datos de entrenamiento y testeo dados los limitantes en la cantidad de datos que se tienen.